Journal of the Korean Data & Information Science Society
2016, **27**(2), 353–362

http://dx.doi.org/10.7465/jkdi.2016.27.2.353
한국데이터정보과학회지

연관성 방향을 고려한 부호 헬링거 측도의 제안

박희창[1]

[1]창원대학교 통계학과

접수 2016년 2월 16일, 수정 2016년 3월 16일, 게재확정 2016년 3월 21일

요 약

데이터 마이닝은 빅 데이터에 내재되어 있는 새로운 법칙이나 잠재되어 있는 지식을 탐색한 후, 이를 근거로 하여 의사결정에 활용하고자 하는 것이다. 위키 백과사전에 의하면 데이터 마이닝 기법 중의 하나인 연관성 규칙은 연관성 평가 기준에 의해 관심 있는 항목들 간에 관련성을 찾아내는 기법으로 많은 연구자들에 의해 연관성 평가를 위한 흥미도 측도들이 개발되어 왔다. 이들 중에서 헬링거 측도는 여러 가지 흥미도 측도들에 비해 많은 장점이 있으나 연관성의 방향을 판단하기가 곤란한 문제를 내포하고 있다. 이 문제를 해결하기 위해 본 논문에서는 부호를 가지는 헬링거 측도를 제안하고 몇 가지 예제를 통하여 유용성을 고찰하였다. 그 결과, 본 논문에서 제안하는 부호 헬링거 측도는 양의 연관성을 가지는 경우에는 양의 값으로 나타나고 음의 연관성을 가지는 경우에는 음의 값을 갖는 것으로 나타났다. 또한 동시발생빈도, 동시 비 발생빈도, 그리고 불일치 빈도가 증가함에 따라 기본적인 연관성 평가 기준들과 부호 헬링거 측도는 증감 여부가 동일한 것을 알 수 있었다.

주요용어: 부호 헬링거 측도, 연관성 규칙, 헬링거 측도, 헬링거 함수, 흥미도 측도.

1. 서론

빅 데이터 기술의 발전은 다변화된 현대사회를 보다 정확하게 예측하고 효율적으로 작동하도록 정보를 제공하는 동시에 과거에는 불가능했던 기술을 가능케 하였으며, 국가차원에서의 사회, 경제, 정치, 문화, 과학기술 등 여러 분야에 활용될 수 있다 (Park과 Pi, 2015). 이러한 빅 데이터 분석 기술 중의 하나가 데이터 마이닝 (data mining)인데, 이는 다양한 형태의 데이터 집합으로부터 잠재되어 있는 지식이나 새로운 법칙과 관계를 탐색한 후, 이를 근거로 의사결정에 활용하고자 하는 것이다 (Park, 2014b).

데이터 마이닝 기법 중에서 많이 활용되고 있는 연관성 규칙 (association rule)은 Agrawal 등 (1993)이 최초로 제안하였으며, 흥미도 측도 (interestingness measure)에 의해 데이터베이스에 포함되어 있는 항목들의 관련성을 도출하는 탐색적 자료 분석 방법으로 응용 분야에서 자주 적용되고 있다 (Park, 2013b). 연관성 규칙의 탐색을 위한 기본적인 흥미도 측도에는 지지도 (support), 신뢰도 (confidence), 그리고 향상도 (lift) 등이 있다. 그런데 이러한 기본적인 측도로는 유용한 연관성 규칙을 찾는 데에는 한계가 있어서 이를 해결하기 위해 많은 연구가 진행되어 왔다 (Silberschatz와 Tuzhilin, 1996; Tan 등, 2002; Ahn과 Kim, 2003; Jin 등, 2011; Park, 2011a; Park, 2011b; Park, 2012a; Park, 2012b; Cho와 Park, 2013; Park, 2013a; Park, 2013b; Lee와 Bae, 2014; Park, 2014a; Park, 2014b; Park, 2015).

바람직한 연관성 규칙을 생성하기 위해 흥미도 측도를 제안한 연구 중에서 Lee와 Bae (2014)는 정보이론을 기반으로 한 헬링거 측도 (Hellinger measures)를 연관성 규칙에 적용하기 위해 제안한 바 있다.

[1] (641-773) 경상남도 창원시 의창구 사림동 9번지, 창원대학교 통계학과, 교수.
E-mail: hcpark@changwon.ac.kr

이 측도는 원래의 헬링거 수식의 제곱형태를 사용함으로써 상대적 정보 내용이 수정된 헬링거 함수에 의해 영향을 받지 않고, 법칙 중요도의 다음 항목인 일반성과의 그 수치 범위 상 합리적인 균형을 이루도록 하였다. 그러나 이 측도는 항상 양의 값을 가지므로 연관성 규칙의 관점에서 볼 때 정의 연관성이 있는지 아니면 부의 연관성이 있는지를 판단하기가 곤란하다. 이 문제를 해결하기 위해 본 논문에서는 연관성 규칙의 방향을 나타낼 수 있는 부호 헬링거 측도 (signed Hellinger measures)를 제안하고자 한다. 논문의 2절에서는 기존의 헬링거 측도를 소개한 후, 부호 헬링거 측도를 제안하고, Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 조건 만족 여부를 증명하며, 3절에서는 예제를 이용하여 이들 측도들을 비교함으로써 제안한 측도의 유용성에 대해 알아본 후, 4절에서 결론을 내리고자 한다.

2. 부호 헬링거 측도

의미 있는 연관성 규칙을 탐색하기 위해서는 기본적인 연관성 평가 기준을 포함하여 각종 흥미도 측도들이 이용된다. Lee와 Bae (2014)는 연관성 규칙을 생성하기 위한 흥미도 측도로 사용하기 위해 Beran (1977)에 의해 소개된 헬링거 함수를 다음과 같이 변형하였다.

$$IC(x \to y) = (\sqrt{p(y|x)} - \sqrt{p(y)})^2 + (\sqrt{1-p(y|x)} - \sqrt{1-p(y)})^2 \tag{2.1}$$

여기서 x는 전항 변수, y는 후항변수, $p(y|x) = P(Y=1|X=1)$, 그리고 $p(y) = P(Y=1)$을 의미한다. 또한 그들은 더 많은 데이터가 연관성 규칙의 조건을 만족하면 더 유용한 법칙이 된다는 연관성 규칙의 일반성을 고려하는 데 있어서 사건의 보편성을 더 정확하게 나타내기 위해 다음과 같이 원래 확률의 제곱근 형태를 사용하였다.

$$G(x \to y) = \sqrt{p(x)} \tag{2.2}$$

여기서 $p(x) = P(X=1)$이다. 이러한 내용을 종합하여 식 (2.1)과 식 (2.2)로부터 연관성 규칙의 최종적인 흥미도 측도로 다음과 같이 헬링거 측도를 제안하였다.

$$H(x \to y) = G(x \to y) \cdot IC(x \to y) = \sqrt{p(x)}\left[(\sqrt{p(y|x)} - \sqrt{p(y)})^2 + (\sqrt{1-p(y|x)} - \sqrt{1-p(y)})^2\right] \tag{2.3}$$

Lee와 Bae (2014)가 제안한 식 (2.3)은 연관성 규칙의 관점에서 살펴보면 양의 신뢰도에서 y의 양의 주변 확률을 뺀 값의 크기와 음의 신뢰도에서 후항의 음의 주변 확률을 뺀 값의 크기를 합해서 이를 x의 양의 주변 확률을 가중치로 고려한 측도이다. 이 측도는 항상 양의 값만을 갖게 되므로 본 논문에서는 이들 두 값의 차이로 나타나는 측도에 지시함수 (indicator function)를 고려하여 다음과 같은 부호 헬링거 측도를 제안하고자 한다.

$$SH(x \to y) = \sqrt{p(x)}\left[I_p(\sqrt{p(y|x)} - \sqrt{p(y)})^2 - I_n(\sqrt{1-p(y|x)} - \sqrt{1-p(y)})^2\right] \tag{2.4}$$

이 식에서 사용한 지시함수는 다음과 같이 정의된다.

$$I_p = \begin{cases} +1, & \text{if } p(y|x) \geq p(y) \\ -1, & \text{if } p(y|x) < p(y) \end{cases}, \quad I_n = \begin{cases} +1, & \text{if } p(\bar{y}|x) \geq p(\bar{y}) \\ -1, & \text{if } p(\bar{y}|x) < p(\bar{y}) \end{cases}$$

여기서 $p(\bar{y}|x) = P(Y=0|X=1)$이고, $p(\bar{y}) = P(Y=0)$이다. 일반적으로 $p(y|x) > p(y)$이면 $p(\bar{y}|x) < p(\bar{y})$가 성립하고, $p(y|x) < p(y)$이면 $p(\bar{y}|x) > p(\bar{y})$가 되므로 식 (2.4)를 연관성 규칙의 기본

적인 평가 기준을 이용하여 정리하면 다음과 같이 기술할 수 있다.

$$SH(x \to y) = \begin{cases} \sqrt{p(x)}[(\sqrt{p(y|x)} - \sqrt{p(y)})^2 + (\sqrt{p(\bar{y}|x)} - \sqrt{p(\bar{y})})^2], & \text{if } p(y|x) > p(y) \\ 0, & \text{if } p(y|x) = p(y) \\ \sqrt{p(x)}[-(\sqrt{p(y|x)} - \sqrt{p(y)})^2 - (\sqrt{p(\bar{y}|x)} - \sqrt{p(\bar{y})})^2], & \text{if } p(y|x) < p(y) \end{cases} \tag{2.5}$$

결국 이 식은 양의 연관성 정도의 크기와 음의 연관성 정도의 크기를 동시에 고려한 동시에 연관성의 방향에 따라 부호가 결정된다. 또한 전항과 후항의 순서가 바뀌면 $p(x)$는 $p(y)$가 되고 $p(y|x)$는 $p(x|y)$가 되므로 일반적으로 $SH(x \to y) \neq SH(y \to x)$가 됨을 알 수 있다. 그러므로 $SH(x \to y)$는 규칙의 원인과 결과를 구분하는 측도라고 할 수 있다.

본 논문에서 제안하는 부호 헬링거 측도 $SH(x \to y)$에 대해 Piatetsky-Shapiro (1991)가 제안한 흥미도 측도의 조건 만족 여부를 알아보고자 한다.

[조건 1] $p(x)$와 $p(y)$가 고정되어 있을 때 $p(xy)$가 증가하면 $SH(x \to y)$는 증가한다. 여기서 $p(xy) = P(X = 1, Y = 1)$을 의미한다.

(증명) 이 특성은 연관성 평가 기준의 중요한 특성 중의 하나로 식 (2.5)의 첫 번째 식을 $p(xy)$에 관해 미분한 후 정리하면 다음과 같다.

$$\frac{\partial SH(x \to y)}{\partial p(xy)} = 1/\sqrt{p(x)}\left[-\sqrt{p(y)/p(y|x)} + \sqrt{p(\bar{y})}/\sqrt{p(\bar{y}|x)}\right] > 0$$

이 경우에는 $p(y|x) > p(y)$이므로 $p(\bar{y}|x) < p(\bar{y})$가 되어 전체적으로 양의 값을 갖게 된다. 따라서 $p(xy)$의 값이 증가함에 따라 증가하는 것을 알 수 있다.

식 (2.5)의 두 번째 식, 즉 $p(y|x) < p(y)$인 경우에 $p(xy)$에 관해 미분하면 다음과 같은 결과를 얻는다.

$$\frac{\partial SH(x \to y)}{\partial p(xy)} = 1/\sqrt{p(x)}\left[\sqrt{p(y)/p(y|x)} - \sqrt{p(\bar{y})}/\sqrt{p(\bar{y}|x)}\right] > 0$$

이 경우에는 $p(y|x) < p(y)$이므로 $p(\bar{y}|x) > p(\bar{y})$가 되어 전체적으로 양의 값을 갖게 된다. 따라서 $p(xy)$의 값이 증가함에 따라 두 번째 수식이 증가하는 것을 알 수 있다.

[조건 2] $SH(x \to y)$는 $p(y)$의 값에 따라 감소한다.

(증명) 식 (2.5)의 첫 번째 식을 $p(y)$에 관해 미분한 후 정리하면 다음과 같다.

$$\frac{\partial SH(x \to y)}{\partial p(y)} = \sqrt{p(x)}\left[-\sqrt{p(y|x)/p(y)} + \sqrt{p(\bar{y}|x)}/\sqrt{p(\bar{y})}\right] < 0$$

여기서 $p(y|x) > p(y)$이고 $p(\bar{y}|x) < p(\bar{y})$이므로 전체적으로 음의 값을 갖게 된다. 따라서 $p(y)$의 값이 증가함에 따라 $SH(x \to y)$가 감소하는 것을 알 수 있다.

식 (2.5)의 두 번째 식을 $p(y)$에 관해 미분하면 다음과 같은 결과를 얻는다.

$$\frac{\partial SH(x \to y)}{\partial p(y)} = \sqrt{p(x)}\left[\sqrt{p(y|x)/p(y)} - \sqrt{p(\bar{y}|x)}/\sqrt{p(\bar{y})}\right] < 0$$

여기서는 $p(y|x) < p(y)$이고 $p(\bar{y}|x) > p(\bar{y})$이므로 전체적으로 음의 값을 갖게 된다. 따라서 $p(y)$의 값이 증가함에 따라 $SH(x \to y)$가 감소하는 것을 알 수 있다.

[조건 3] 전항 X와 후항 Y가 서로 독립이면 $SH(x \to y) = 0$이 된다.

(증명) X와 Y가 서로 독립이면 식 (2.5)의 각 식에서 괄호 안의 첫 번째 식과 두 번째 식이 모두 0이 되므로 $SH(x \to y) = 0$이 됨을 알 수 있다.

3. 적용 예제

이 절에서는 예제를 이용하여 부호 헬링거 측도인 $SH(x \to y)$의 유용성을 고찰하고자 한다. 이를 위해 먼저 다음과 같은 사례를 통하여 기존의 $H(x \to y)$와 $SH(x \to y)$의 측도를 비교해보기로 한다.

Table 3.1 Case 1

		Y		Total
		1	0	
X	1	0.03	0.07	0.10
	0	0.17	0.73	0.90
Total		0.20	0.80	1.00

Table 3.1에서 X와 Y에 대한 주변 비율은 각각 $p(x) = 0.1$, $p(y) = 0.2$이고, 양의 신뢰도는 $P(y|x) = 0.3$, 음의 신뢰도는 $P(y|\bar{x}) = 0.19$이며, 향상도는 $P(y|x)/p(y) = 1.5$로 나타나서 전체적으로 양의 연관성이 존재한다고 볼 수 있다. 이 사례에서 $H(x \to y) = 0.0042$이고, $p(y|x) > p(y)$이므로 $SH(x \to y)$는 $H(x \to y)$와 동일한 양의 값을 갖게 된다.

다음으로는 음의 연관성인 경우에 대해 헬링거 측도들을 비교해보기 위해 Table 3.2의 사례를 고려하고자 한다.

Table 3.2 Case 2

		Y		Total
		1	0	
X	1	0.03	0.07	0.10
	0	0.57	0.33	0.90
Total		0.60	0.40	1.00

Table 3.2에서 $p(x) = 0.1$, $p(y) = 0.6$이고, 양의 신뢰도는 $P(y|x) = 0.3$, 음의 신뢰도는 $P(y|\bar{x}) = 0.63$이며, 향상도는 $P(y|x)/p(y) = 0.5$로 나타나서 항목 X와 Y간에는 음의 연관성이 존재한다. 이 사례에서 $H(x \to y) = 0.0295$이고, $p(y|x) < p(y)$이므로 $SH(x \to y) = -0.0295$가 되어 $H(x \to y)$와 부호가 반대인 값을 갖는다.

Table 3.1과 Table 3.2의 두 표를 비교해보면 기존의 $H(x \to y)$측도는 항상 양의 값을 갖게 되어 연관성의 방향을 나타내주지 못하는 반면에 본 논문에서 제안하는 측도인 $SH(x \to y)$는 양의 연관성을 가지는 경우에는 양의 값으로 나타나고 음의 연관성을 가지는 경우에는 음의 값을 갖는 것으로 나타나므로 이 측도는 $H(x \to y)$와 동일한 값을 가지는 동시에 연관성의 방향을 나타내는 측도라고 할 수 있다.

측도 $SH(x \to y)$의 유용성을 좀 더 자세히 탐색하기 위해 Park (2011)에서 고려한 모의실험 데이터를 활용하고자 한다. 먼저 동시발생빈도가 증가하는 경우에 $SH(x \to y)$의 변화하는 양상을 알아보기 위해 Table 3.3의 데이터를 이용한다.

Table 3.3 Simulation data (1)

		Y		Total
		1	0	
X	1	a	$50 - a$	50
	0	$30 - a$	$a + 20$	50
Total		30	70	100

Table 3.3에서 항목 X와 Y의 주변 확률은 각각 $p(x) = 0.5$과 $p(y) = 0.3$이다. Table 3.3으로부터 지지도 $supp = p(xy)$, 신뢰도 $conf_+ = p(y|x)$, $conf_- = p(\bar{y}|x)$, 향상도 $lift = p(y|x)/p(y)$, 그리고 부호 헬링거 측도 SH를 계산하면 Table 3.4의 결과를 얻을 수 있다. 여기서 a는 동시 발생 빈도 $P(X = 1, Y = 1)$, d는 동시 비 발생빈도 $P(X = 0, Y = 0)$, b와 c는 각각 불일치 빈도 $P(X = 1, Y = 0)$과 $P(X = 0, Y = 1)$을 의미한다. 또한 이를 그림으로 나타내면 Figure 3.1과 같다. 그림에서 $conf_1$은 $conf_+$를 의미한다.

이들 표와 그림에서 보는 바와 같이 동시발생빈도 a가 증가함에 따라 지지도, $conf_+$, 향상도, 그리고 측도 SH가 증가하는 반면에 $conf_-$는 감소하고 있다. 따라서 기본적인 연관성 평가 기준들과 SH는 증감 여부가 동일한 것을 알 수 있다. 또한 향상도의 값이 1보다 작은 경우에는 SH가 음의 값으로 나타났고, 향상도의 값이 1보다 큰 경우에는 SH가 양의 값으로 나타났으므로 SH는 연관성의 방향을 잘 표현하는 것을 알 수 있다.

Table 3.4 Outputs of signed Hellinger measure by simulation data (1)

a	b	c	d	$supp$	$conf_+$	$conf_-$	$lift$	SH
1	49	29	21	0.010	0.020	0.980	0.067	-0.1333
2	48	28	22	0.020	0.040	0.960	0.133	-0.1000
3	47	27	23	0.030	0.060	0.940	0.200	-0.0773
4	46	26	24	0.040	0.080	0.920	0.267	-0.0602
5	45	25	25	0.050	0.100	0.900	0.333	-0.0468
6	44	24	26	0.060	0.120	0.880	0.400	-0.0359
7	43	23	27	0.070	0.140	0.860	0.467	-0.0271
8	42	22	28	0.080	0.160	0.840	0.533	-0.0199
9	41	21	29	0.090	0.180	0.820	0.600	-0.0141
10	40	20	30	0.100	0.200	0.800	0.667	-0.0095
11	39	19	31	0.110	0.220	0.780	0.733	-0.0059
12	38	18	32	0.120	0.240	0.760	0.800	-0.0032
13	37	17	33	0.130	0.260	0.740	0.867	-0.0014
14	36	16	34	0.140	0.280	0.720	0.933	-0.0003
15	35	15	35	0.150	0.300	0.700	1.000	0.0000
16	34	14	36	0.160	0.320	0.680	1.067	0.0003
17	33	13	37	0.170	0.340	0.660	1.133	0.0013
18	32	12	38	0.180	0.360	0.640	1.200	0.0029
19	31	11	39	0.190	0.380	0.620	1.267	0.0051
20	30	10	40	0.200	0.400	0.600	1.333	0.0078
21	29	9	41	0.210	0.420	0.580	1.400	0.0111
22	28	8	42	0.220	0.440	0.560	1.467	0.0150
23	27	7	43	0.230	0.460	0.540	1.533	0.0194
24	26	6	44	0.240	0.480	0.520	1.600	0.0243
25	25	5	45	0.250	0.500	0.500	1.667	0.0298
26	24	4	46	0.260	0.520	0.480	1.733	0.0359
27	23	3	47	0.270	0.540	0.460	1.800	0.0425
28	22	2	48	0.280	0.560	0.440	1.867	0.0497
29	21	1	49	0.290	0.580	0.420	1.933	0.0575

이를 좀 더 구체적으로 기술하기 위해 ($a = 7$, $b = 43$, $c = 23$, $d = 27$)인 경우와 ($a = 21$, $b = 29$, $c = 9$, $d = 41$)인 경우를 비교해보면 각 경우에 대해 항목 X에 대한 발생 비율이 0.5이므로 $\sqrt{p(x)} = 0.7071$이 되며, 전자의 경우에는 $\sqrt{p(y|x)} - \sqrt{p(y)}$와 $\sqrt{p(\bar{y}|x)} - \sqrt{p(\bar{y})}$의 값이 각각 -0.1736과 0.0907이 되고, 후자의 경우에는 각각 0.1004와 -0.0751이 된다. 즉, 전자에서는 양의 연관성을 나타내는 값이 음으로 계산되었고, 음의 연관성을 나타내는 값이 양으로 계산된 반면에 후자에서는 양의 연관성을 나타내는 값이 양으로 얻어졌고 음의 연관성을 나타내는 값이 음으로 얻어졌다. 이에 따라 SH의 값이 전자의 경우에는 음의 값인 -0.0271로, 후자의 경우에는 양의 값인 0.0111로 나타나서 연관성 평가 기준으로 SH를 이용할 경우 크기에 의해 연관성의 정도를 알 수 있고 부호에 의해 연관성을 방향을 파악할 수 있다.

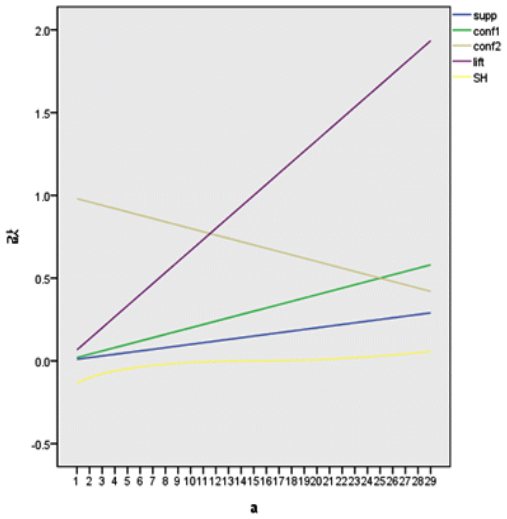

Figure 3.1 Outputs of signed Hellinger measure by simulation data (1)

다음으로는 Table 3.5의 교차표를 이용하여 불일치빈도 b의 증감에 따른 $SH(x \to y)$의 변화하는 양상을 알아보고자 한다.

Table 3.5 Simulation data (2)

		Y		Total
		1	0	
X	1	$50-b$	b	50
	0	$b+20$	$30-b$	50
Total		70	30	100

Table 3.5에서 항목 X와 Y의 주변 확률은 각각 $p(x) = 0.5$과 $p(y) = 0.7$이다. Table 3.5로부터 $supp$, $conf_+$, $conf_-$, $lift$, 그리고 측도 SH를 계산하면 Table 3.6 및 Figure 3.2의 결과를 얻을 수 있다. 그림에서 $conf_2$는 $conf_-$를 의미한다. 이 표에서 보는 바와 같이 불일치빈도 b가 증가함에 따라 지지도, $conf_+$, 향상도, 그리고 SH가 감소하는 반면에 $conf_-$는 증가하고 있다. 이 표에서도 측도 SH의 증감 여부가 기본적인 연관성 평가 기준들과 동일한 것을 알 수 있다. 또한 향상도의 값이 1보다 큰 경우에는 SH가 양의 값으로 계산되었고, 향상도의 값이 1보다 작은 경우에는 SH가 양의 값으로 계산되었으므로 SH는 연관성의 방향을 적절하게 나타내고 있다.

이에 대해 구체적으로 알아보기 위해 ($a = 45$, $b = 5$, $c = 25$, $d = 25$)인 경우와 ($a = 28$, $b = 22$, $c = 42$, $d = 8$)인 경우를 비교해보면 각 경우에 대해 $\sqrt{p(x)} = 0.7071$이 되며, 전자의 경우에는 $\sqrt{p(y|x)} - \sqrt{p(y)}$와 $\sqrt{p(\bar{y}|x)} - \sqrt{p(\bar{y})}$의 값이 각각 0.1120과 -0.2315가 되고, 후자의 경우에는 각각 -0.0883과 0.1156이 된다. 즉, 전자에서는 양의 연관성을 나타내는 수식은 양의 값으로 나타났고, 음의 연관성을 나타내는 수식은 음의 값으로 나타난 반면에 후자에서는 양의 연관성을 나타내는 수식은 음의 값으로 계산되었고 음의 연관성을 나타내는 수식은 음의 값으로 계산되었다. 따라서 SH의 값이 전자의 경우에는 양의 값인 0.0468로, 후자의 경우에는 음의 값인 -0.0150으로 나타나서 이 경우에도 SH는 그 부호에 의해 연관성을 방향을 파악할 수 있는 것으로 확인이 되었다.

Table 3.6 Outputs of signed Hellinger measure by simulation data (2)

a	b	c	d	$supp$	$conf_+$	$conf_-$	$lift$	SH
50	0	20	30	0.500	1.000	0.000	1.429	0.2310
49	1	21	29	0.490	0.980	0.020	1.400	0.1333
48	2	22	28	0.480	0.960	0.040	1.371	0.1000
47	3	23	27	0.470	0.940	0.060	1.343	0.0773
46	4	24	26	0.460	0.920	0.080	1.314	0.0602
45	5	25	25	0.450	0.900	0.100	1.286	0.0468
44	6	26	24	0.440	0.880	0.120	1.257	0.0359
43	7	27	23	0.430	0.860	0.140	1.229	0.0271
42	8	28	22	0.420	0.840	0.160	1.200	0.0199
41	9	29	21	0.410	0.820	0.180	1.171	0.0141
40	10	30	20	0.400	0.800	0.200	1.143	0.0095
39	11	31	19	0.390	0.780	0.220	1.114	0.0059
38	12	32	18	0.380	0.760	0.240	1.086	0.0032
37	13	33	17	0.370	0.740	0.260	1.057	0.0014
36	14	34	16	0.360	0.720	0.280	1.029	0.0003
35	15	35	15	0.350	0.700	0.300	1.000	0.0000
34	16	36	14	0.340	0.680	0.320	0.971	-0.0003
33	17	37	13	0.330	0.660	0.340	0.943	-0.0013
32	18	38	12	0.320	0.640	0.360	0.914	-0.0029
31	19	39	11	0.310	0.620	0.380	0.886	-0.0051
30	20	40	10	0.300	0.600	0.400	0.857	-0.0078
29	21	41	9	0.290	0.580	0.420	0.829	-0.0111
28	22	42	8	0.280	0.560	0.440	0.800	-0.0150
27	23	43	7	0.270	0.540	0.460	0.771	-0.0194
26	24	44	6	0.260	0.520	0.480	0.743	-0.0243
25	25	45	5	0.250	0.500	0.500	0.714	-0.0298
24	26	46	4	0.240	0.480	0.520	0.686	-0.0359
23	27	47	3	0.230	0.460	0.540	0.657	-0.0425
22	28	48	2	0.220	0.440	0.560	0.629	-0.0497
21	29	49	1	0.210	0.420	0.580	0.600	-0.0575
20	30	50	0	0.200	0.400	0.600	0.571	-0.0659

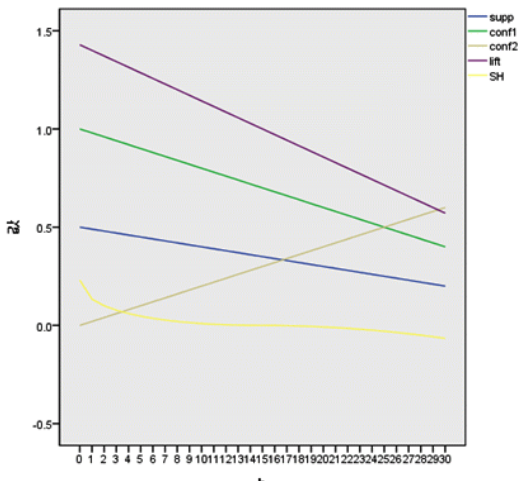

Figure 3.2 Outputs of signed Hellinger measure by simulation data (2)

또 다른 불일치빈도 c및 동시 비 발생빈도 d의 변화에 따라 $SH(x \to y)$가 변화하는 양상을 살펴보았는데, 전자의 경우에는 Table 3.4와 유사한 결과를 얻었으며, 후자의 경우에는 Table 3.6과 유사한 결과를 얻게 되었다.

4. 결론

오늘날 빅 데이터가 신 성장 동력으로 부상하고 있는 가운데 이를 분석하여 유용한 정보를 찾아내 주는 데이터마이닝 기술이 주목받고 있다 (Park, 2013a). 본 논문에서는 Lee와 Bae (2014)가 제안한 헬링거 측도에 대해 연관성의 방향을 나타낼 수 있는 부호 헬링거 측도를 개발하여 흥미도 측도의 조건 충족 여부를 점검하였다. 또한 예제를 이용하여 제안한 측도의 유용성을 살펴보았다. 먼저 정의 연관성 규칙과 부의 연관성 규칙을 구분할 수 있는지에 대해 사례를 통해 검토한 결과, 기존의 측도는 모두 양의 값으로 계산되어서 연관성의 방향을 파악할 수 없었으나 부호 헬링거 측도 SH는 정의 경우에는 양의 값으로, 그리고 부의 경우에는 음의 값으로 나타나서 연관성의 방향을 파악할 수 있었다. 또한 제안한 측도의 유용성을 좀 더 자세히 탐색하기 위해 동시발생빈도가 증가하는 경우와 불일치 빈도가 증가하는 경우에 대해 변화하는 양상을 탐색하였다. 그 결과, 동시발생빈도가 증가함에 따라 지지도, $conf_+$, 향상도, 그리고 측도 SH가 증가하는 반면에 $conf_-$는 감소하였다. 따라서 기본적인 연관성 평가 기준들과 SH는 증감 여부가 동일한 것을 알 수 있다. 불일치빈도가 증가함에 따라 지지도, $conf_+$, 향상도, 그리고 SH가 감소하는 반면에 $conf_-$는 증가하였다. 이 경우에도 측도 SH의 증감 여부가 기본적인 연관성 평가 기준들과 동일한 것을 알 수 있었다. 두 경우 모두 향상도의 값이 1보다 작은 경우에는 SH가 음의 값으로 나타났고, 향상도의 값이 1보다 큰 경우에는 SH가 양의 값으로 나타났으므로 SH는 그 부호에 의해 연관성을 방향을 파악할 수 있는 것으로 확인이 되었다.

References

Agrawal, R., Imielinski, R. and Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD Conference on Management of Data*, Association for Computing Machinery, New York, USA, 207-216.

Ahn, K. and Kim, S. (2003). A new interstingness measure in association rules mining. *Journal of the Korean Institute of Industrial Engineers*, **29**, 41-48.

Beran, R. J. (1977). Minimum hellinger distances for parametric models. *Annals of Statistics*, **5**, 445-463.

Cho, K. H. and Park, H. C. (2013). A study of Gyungnam's social indicator survey using data mining. *Journal of the Korean Data Analysis Society*, **15**, 2489-2497.

Lee, C. H. and Bae, J. H. (2014). A new importance measure of association rules using information theory. *Journal of rhe Korea Information Processing Society Transactions on Software and Data Engineering*, **3**, 37-42.

Jin, D. S., Kang, C., Kim, K. K. and Choi, S. B. (2011). CRM on travel agency using association rules. *Journal of the Korean Data Analysis Society*, **13**, 2945-2952.

Park, H. C. (2011a). Association rule ranking function by decreased lift influence. *Journal of the Korean Data & Information Science Society*, **22**, 179-188.

Park, H. C. (2011b). The proposition of attributably pure confidence in association rule mining. *Journal of the Korean Data & Information Science Society*, **22**, 235-243.

Park, H. C. (2012a). Negatively attributable and pure confidence for generation of negative association rules. *Journal of the Korean Data & Information Science Society*, **23**, 707-716.

Park, H. C. (2012b). Exploration of PIM based similarity measures as association rule thresholds. *Journal of the Korean Data & Information Science Society*, **23**, 1127-1135.

Park, H. C. (2013a). The proposition of c ompared and a ttributably pure confidence in association rule mining. *Journal of the Korean Data & Information Science Society*, **24**, 523-532.

Park, H. C. (2013b). Proposition of causal association rule thresholds. *Journal of the Korean Data & Information Science Society*, **24**, 1189-1197.

Park, H. C. (2014a). Comparison of cosine family similarity measures in the aspect of association rule. *Journal of the Korean Data Analysis Society*, **16**, 729-737.

Park, H. C. (2014b). Comparison of confidence measures useful for classification model building. *Journal of the Korean Data & Information Science Society*, **25**, 1-7.

Park, H. C. (2015). A study on the ordering of PIM family similarity measures without marginal probability. *Journal of the Korean Data & Information Science Society*, **26**, 367-376.

Park, J. H. and Pi, S. Y. (2015). A study on wt-algorithm for effective reduction of association rules. *Journal of the Korea Industrial Information Systems Research*, **20**, 61-69.

Piatetsky-Shapiro, G. (1991). *Discovery, analysis and presentation of strong rules*, Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge MA, USA, 229-248.

Silberschatz, A. and Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge Data Engineering*, **8**, 970-974.

Tan, P. N., Kumar, V. and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, USA, 32-41.

Journal of the Korean Data & Information Science Society
2016, **27**(2), 353–362

http://dx.doi.org/10.7465/jkdi.2016.27.2.353
한국데이터정보과학회지

Signed Hellinger measure for directional association

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 16 February 2016, revised 16 March 2016, accepted 21 March 2016

Abstract

By Wikipedia, data mining is the process of discovering patterns in a big data set involving methods at the intersection of association rule, decision tree, clustering, artificial intelligence, machine learning. and database systems. Association rule is a method for discovering interesting relations between items in large transactions by interestingness measures. Association rule interestingness measures play a major role within a knowledge discovery process in databases, and have been developed by many researchers. Among them, the Hellinger measure is a good association threshold considering the information content and the generality of a rule. But it has the drawback that it can not determine the direction of the association. In this paper we proposed a signed Hellinger measure to be able to interpret operationally, and we checked three conditions of association threshold. Furthermore, we investigated some aspects through a few examples. The results showed that the signed Hellinger measure was better than the Hellinger measure because the signed one was able to estimate the right direction of association.

Keywords: Association rule, Hellinger divergence, Hellinger measure, interestingness measure, signed Hellinger measure.

[1] Professor, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea. E-mail: hcpark@changwon.ac.kr